Homework Set #3:

1. When a thermal oxide of thickness 0.50 μm is grown on a silicon wafer using either wet or dry oxidation, what thickness of the substrate is consumed? What is the apparent color of this oxide layer when viewed normally from above? Where are the top and bottom surfaces of the resulting oxide layer located relative to the original surface of the wafer? Does any of this depend on crystal orientation (*i.e.*, will results differ on [100] and [111] wafers)? Why or why not? (Assume a density of elemental silicon as 2.33 g/cm^3 and a density of amorphous silicon dioxide, *i.e.*, fused quartz or quartz glass, as 2.27 g/cm^3.)

The density of silicon is 2.33 g/cm^3 and the density of amorphous silicon dioxide is 2.27 g/cm^3. Thus, dividing both of these values by their formula weights, one finds that the formula weight density of silicon is 0.0829 FW/cm^3 *and the formula weight density of amorphous silicon dioxide is 0.0378* FW/cm^3. *Thus, the ratio of the formula weight density of amorphous silicon dioxide to silicon is 0.4555. Therefore, for a given silicon dioxide film of arbitrary thickness, x_o:*

$$\#\text{FW's SiO}_2 = 0.0378\ \text{FW/cm}^3 \times \text{wafer area} \times x_o$$

Similarly, the number of formula weights of silicon consumed is given by:

$$\#\text{FW's Si} = 0.0829\ \text{FW/cm}^3 \times \text{wafer area} \times x_{Si}$$

where, x_{Si} is the thickness of silicon consumed. Clearly, the stoichometry of the growth reaction requires:

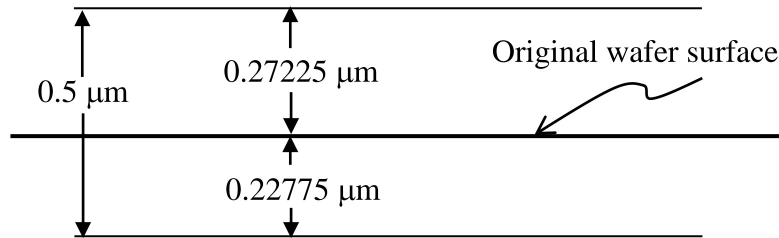
$$\#\text{FW's Si} = \#\text{FW's SiO}_2$$

Thus, one finds:

$$\frac{x_{Si}}{x_o} = \frac{0.0378\ \text{FW/cm}^3}{0.0829\ \text{FW/cm}^3} = 0.4555$$

Therefore, for any thermal oxidation process the thickness of silicon consumed is about 45.5% the thickness of amorphous silicon dioxide grown. Thus, for a 0.5 μm *thick thermal oxide, 0.22775* μm *of the silicon substrate must be consumed.*

The relative positions of top and bottom surfaces of the oxide layer relative to the original wafer surface are illustrated by the following figure



Although thermal oxide is very transparent at optical and ultraviolet wavelengths, diffraction imparts an apparent color. For 0.50 μm *as indicated by the table in the notes, this color is blue-green, which tends toward blue for slightly thinner oxide and toward a very prominent green color for slightly thicker oxide.*

Thermal oxide thickness relative to the quantity of silicon consumed is independent of wafer orientation since the volume concentration of silicon atoms is independent of wafer orientation Of course, the growth rate is different on [111] silicon versus [100] silicon and can be understood as a consequence of differing areal concentration of silicon atoms associated with different crystal planes.

2. Using the Arrhenius data (activation energies and pre-exponential coefficients) given in the class notes, calculate the linear and parabolic rate constants for dry oxidation of [111] and [100] silicon at 700, 900, and 1100C. Which wafer orientation oxidizes more in a given time and why? If it is necessary to use 700C for this oxidation process in order to maintain an acceptable thermal budget for the overall process, is there any practical way to increase the rate by a factor of fifteen without increasing the temperature?

A process which obeys an Arrhenius law is a thermally activated process. The rate constants of such a process have the general form:

$$\kappa = \kappa_0 \exp\left(-\frac{E_a}{kT}\right)$$

For oxidation, κ *can be either the linear or parabolic rate constant,* κ_0 *is a the "pre-exponential factor" and corresponds to the rate constant at "infinite temperature", that is to say, the rate when the average thermal energy of the system greatly exceeds the activation energy. Of course,* E_a *is activation energy, which physically can be thought of as deriving from an energy barrier existing between so-called reactant and product states.*

First of all, ordinary temperature must be converted to absolute temperature. Thus, one finds:

700C = 973K; 900C = 1173K; 1100C = 1373K

From the data given in the notes, it follows that the linear rate constant, B/A, *for dry oxidation is given on a [100] silicon surface by:*

$$\frac{B}{A} = 1.03(10^3)\ \mu\text{m/sec} \times \exp\left(-\frac{2.0\ \text{eV}}{kT}\right)$$

and on a [111] silicon surface by:

$$\frac{B}{A} = 1.73(10^3)\ \mu\text{m/sec} \times \exp\left(-\frac{2.0\ \text{eV}}{kT}\right)$$

Again, it is clear that these two expressions differ only by the pre-exponential factor. Naturally, this is to be expected since the difference in the rates is due to the difference in the effective concentration of silicon atoms on the two types of surfaces. Obviously, the energetics of the oxidation reaction remain the same for both [100] and [111] silicon surfaces. Therefore, one would not expect the activation energies to be different on [100] and [111] surfaces. This is confirmed experimentally by the observation that both rate constants have the same slope on an Arrhenius plot (i.e., a plot of the logarithm of the rate constant vs reciprocal temperature). The parabolic rate constant, B, for dry oxidation is given on [100] and [111] by the same expression:

$$B = 0.214\ \mu\text{m}^2/\text{sec} \times \exp\left(-\frac{1.23\ \text{eV}}{kT}\right)$$

Once again, this makes physical sense since in the parabolic regime film growth is limited by oxidant diffusion through the oxide layer. Clearly, since quartz glass, i.e., thermal oxide, is amorphous and its structure is independent of any original silicon crystal structure, this process is independent of the orientation of the underlying silicon surface and, thus, is same for oxidation of either [100] or [111] silicon surfaces.

Using a value of $8.6173(10^{-5})\ \text{eV}/^\circ\text{K}$ *for Boltzmann's constant, k, one finds:*

At 700C*:*
$B/A = 4.5040(10^{-8})\ \mu\text{m/sec}$ *on [100]*
$B/A = 7.5650(10^{-8})\ \mu\text{m/sec}$ *on [111]*
$B = 9.1093(10^{-8})\ \mu\text{m}^2/\text{sec}$ *on [100] and [111]*

At 900C*:*
$B/A = 2.6296(10^{-6})\ \mu\text{m/sec}$ *on [100]*
$B/A = 4.4166(10^{-6})\ \mu\text{m/sec}$ *on [111]*
$B = 1.1111(10^{-6})\ \mu\text{m}^2/\text{sec}$ *on [100] and [111]*

At 1100C*:*
$B/A = 4.6945(10^{-5})$ μm/sec *on [100]*
$B/A = 7.8849(10^{-5})$ μm/sec *on [111]*
$B = 6.5395(10^{-6})$ μm^2/sec *on [100] and [111]*

As observed above, [111] oxidizes faster than [100] because the surface concentration of silicon atoms is higher on a [111] surface.

Of course, if process temperature cannot be changed, then one must resort to some other means of increasing the oxidation rate. One possibility is the use of wet oxidation instead of dry oxidation since the intrinsic rate constants for wet oxidation are larger. However, wet oxidation may result in poorer interfacial quality for the oxide, but if this is not an issue then wet oxidation may be used. Alternatively, oxidation rate constants scale directly with oxidant pressure, therefore, carrying out the oxidation at 15 atmospheres instead of one atmosphere should give the desired result. However, a possible problem with such a process change might be increased generation of oxidation induced defects (typically extrinsic stacking faults) due to the increased growth rate.

3. Suppose you are responsible for a dry oxidation process designed to grow 120 nm of oxide on [100] wafers:

a. Suppose that process integration issues require that a temperature of no higher than 900C must be used. What is the time you will use to obtain this thickness? What is the time if steam can be used instead?

Both dry and wet oxidation are governed by the Deal-Grove growth law:

$$x^2 + Ax = B(t + t_0)$$

Of course, A and B are the rate constants. This expression can be rearranged as follows:

$$t = \frac{x^2}{B} + \frac{x}{B/A} - t_0$$

Here, t is the time needed to grow an oxide layer of thickness, x, and t_0 is an initial condition, i.e., either an actual or a hypothetical time required to grow any pre-existing oxide film under designated process conditions. In this form, the utility of the designation of B as the parabolic rate constant and B/A as the linear rate constant is obvious. Following the same procedure as in the previous problem, one finds that at 900C *(i.e.,* 1173K*) on [100] silicon wafers:*

$B/A = 2.6296(10^{-6})$ μm/sec
$B = 1.1111(10^{-6})$ μm^2/sec

Now, for dry oxidation, a fictitious initial thickness of 0.020 μm *is assumed. The corresponding initial condition,* t_0*, is obtained directly from the Deal-Grove law:*

$$t_0 = \frac{x_0^{\,2}}{B} + \frac{x_0}{B/A} = \frac{(0.020\ \mu\text{m})^2}{1.11(10^{-6})\ \mu\text{m}^2/\text{sec}} + \frac{0.020\ \mu\text{m}}{2.63(10^{-6})\ \mu\text{m/sec}} = 7965\ \text{sec}$$

One now substitutes this result to obtain:

$$t = \frac{(0.120\ \mu\text{m})^2}{1.11(10^{-6})\ \mu\text{m}^2/\text{sec}} + \frac{0.120\ \mu\text{m}}{2.63(10^{-6})\ \mu\text{m/sec}} - 7965\ \text{sec}$$

$$t = 50630\ \text{sec} = 14.06\ \text{hr}$$

If steam was used instead, the procedure remains very similar. Of course, one must use the rate constants appropriate to wet oxidation. It follows from the data given in the class notes that for a [100] silicon surface:

$$\frac{B}{A} = 2.70(10^4)\ \mu\text{m/sec} \times \exp\left(-\frac{2.05\ \text{eV}}{kT}\right)$$

$$B = 0.107\ \mu\text{m}^2/\text{sec} \times \exp\left(-\frac{0.79\ \text{eV}}{kT}\right)$$

Thus, at 900C*, one obtains:*

$B/A = 4.2032(10^{-5})$ μm/sec
$B = 4.3168(10^{-6})$ μm²/sec

For wet oxidation, no initial rapid oxidation occurs, thus, no initial fictitious oxide thickness is needed and t_0 *can be taken to be zero. Therefore:*

$$t = \frac{(0.120\ \mu\text{m})^2}{4.31(10^{-5})\ \mu\text{m}^2/\text{sec}} + \frac{0.120\ \mu\text{m}}{4.20(10^{-5})\ \mu\text{m/sec}}$$

$$t = 3189\ \text{sec} = 0.886\ \text{hr}$$

Clearly, wet oxidation is much faster than dry oxidation. In general, a processing time of 14.06 hours is unacceptably long in a modern manufacturing environment. This result clearly illustrates the "throughput" advantage of wet oxidation over dry oxidation. However, dry oxidation is advantageous for fabrication of very thin oxides since the interface state density for dry oxides is generally lower.

b. Suppose that the furnace aborts during the process you designed in part **a**. You remove the wafers, measure the oxide thickness and find that there is 100 nm of oxide. Once the furnace is repaired, what do you do now? If you can "save" these wafers, how do you do it?

This is a realistic situation that might confront a process engineer. Of course, it may be better to "scrap" the wafers for such a case of non-standard processing. However, if the oxide layer is not that critical, e.g., not a gate oxide, it may be convenient and economical to "rework". In this case, once the furnace is repaired and tested, one can finish the oxidation by calculating the time needed to add the required amount of oxide. Thus, if x_i *is an initial amount of oxide present on the wafer, the Deal-Grove law takes the form:*

$$t = \frac{x^2}{B} + \frac{x}{B/A} - \frac{x_i^2}{B} - \frac{x_i}{B/A}$$

Clearly, the time required to add the additional oxide is just the difference in the time it took to grow the 100 nm and the time required to grow the full 120 nm oxide layer. (It turns out that in this form, a fictitious initial thickness is not needed explicitly provided that x_i *is more than 20-25 nm in the case of a dry oxidation.) Thus,* x_i *is the 100 nm of oxide grown during the aborted furnace cycle. Of course, the rate constants are the same as before.*

Accordingly, it was found in part **a**, *that growth of 120 nm of oxide requires a time of 14.06 hrs. Thus, it follows that the time, required to grow 100 nm of oxide, i.e., the time until the abort occurred which provides the initial condition for the regrowth, is given by a corresponding calculation:*

$$t_{abort} = \frac{(0.100\ \mu\text{m})^2}{1.11(10^{-6})\ \mu\text{m}^2/\text{sec}} + \frac{0.100\ \mu\text{m}}{2.63(10^{-6})\ \mu\text{m/sec}} - 7965\ \text{sec}$$

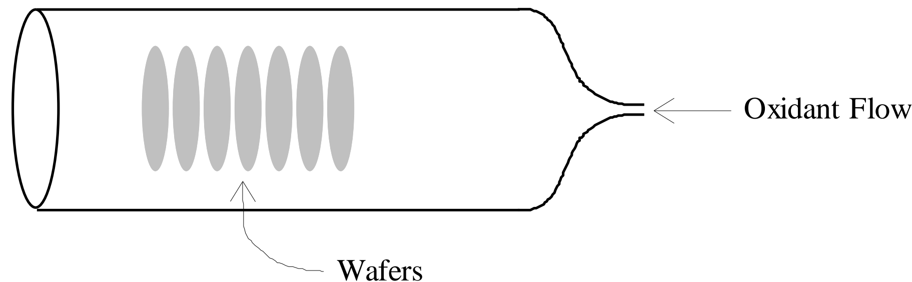
$$t_{abort} = 39067\ \text{sec} = 10.85\ \text{hr}$$

(Here, initial rapid oxidation must be accounted for because it is actual oxidation time that is being calculated.) Thus, the furnace must have been running for just over seven hours before aborting. The time needed to complete the oxidation is just the difference of the time needed to grow 120 nm and the time needed to grow 100 nm. Therefore, the additional time, t, is:

$$t = 11563\ \text{sec} = 3.212\ \text{hr}$$

Thus, the wafers need to go back into the furnace for about three and a quarter hours to complete the oxidation.

c. During processing in either a vertical or horizontal quartz tube furnace wafers are generally held at the edges by a quartz holder (or "boat"). The wafers are conventionally held perpendicular to the gas stream as shown below:



Your first attempt at designing the process described in part **a** results in wafers with unacceptable oxide uniformity. In particular, they exhibit a gradient from top to bottom with the oxide at the top being thicker than at the bottom. What might be causing this gradient and how would you go about fixing it? Is it due to oxidant depletion within the furnace or something else? Why or why not?

The oxide thickness gradient observed on the wafers is almost certainly due to temperature non-uniformity. This can arise by not allowing the wafers to come to thermal equilibrium in the furnace before starting the oxidation, i.e., the "soak" step is too short, or by too low a gas flow which causes the furnace to be unevenly heated, i.e., convection within the furnace is not damped out. Other more pathological causes due to leaks or "aspiration" may also cause such non-uniformity.

In any case, it is clear that the wafers are getting hotter on the top than on the bottom. This may be solved by increasing gas flow and/or increasing thermal soak before processing. It is not due to oxidant depletion since, as indicated by the Deal-Grove kinetic model of oxidation, an oxidation process never enters a mass transport limited regime.

4. Consider an aluminum "gate" MOS capacitor fabricated on a uniformly doped *p*-type substrate covered with a uniform layer of thermally grown silicon dioxide. Suppose that the net acceptor doping concentration is 5×10^{15} atoms/cm^3, that the gate electrode is circular with diameter 100 μm, and that the oxide thickness is 50 nm.

a. Calculate minimum and maximum absolute capacitance one expects to obtain for a high frequency CV measurement made at 300K. (Note: ε_{ox}=0.34 pF/cm and ε_s=1.04 pF/cm)

The oxide capacitance per unit area is easily obtained as follows:

$$C_{ox} = \frac{\varepsilon_{ox}}{x_o} = \frac{0.34\,\text{pF/cm}}{5(10^{-6})\,\text{cm}} = 68.00\,\text{nF/cm}^2$$

The maximum depletion width is calculated from the standard formula:

$$x_d^{\max} = \sqrt{\frac{4\varepsilon_s kT}{q^2 N_A}\ln\left(\frac{N_A}{n_i}\right)} =$$

$$\sqrt{\frac{4(1.04\,\text{pF/cm})(8.6173(10^{-5})\,\text{eV/°K})(300\text{°K})}{(1.602(10^{-19})\,\text{C})(1\,\text{eV/V})(5(10^{15})\,\text{cm}^{-3})}\ln\left(\frac{5(10^{15})\,\text{cm}^{-3}}{1.042(10^{10})\,\text{cm}^{-3}}\right)} = 0.4191\,\mu\text{m}$$

Some explanation of units rationalization is necessary here. Clearly, in the expression for maximum depletion width, the square of q, the fundamental charge, appears explicitly in the denominator within the radical. It turns out that for rationalization of units, it is advantageous to express each of the two factors of q in a different set of units. Thus, one factor of q is expressed in ordinary cgs units, i.e., Coulombs ($1.602(10^{-19})$ C*). In contrast, the other factor of q is expressed in atomic units, i.e., electron volts per volt. Clearly, since energy is the product of charge and potential,* eV/V *must be a unit of charge. In particular, one* eV *is defined as the amount of energy that a particle, e.g., an electron, having one fundamental unit of charge, i.e., q, gains when "falling through" a potential drop of one volt. Hence it follows that q must be precisely equal to one* eV/V. *(Therefore,* 1 eV/V *is exactly equivalent to* $1.602(10^{-19})$ C.*) Furthermore, since one Coulomb is defined as the product of one Farad and one volt, this substitution "trick" allows all energy and electrical units to be cancelled out leaving only units of length (as is desired). Thus, the maximum depletion capacitance per unit area now follows immediately:*

$$C_s = \frac{\varepsilon_s}{x_d^{\max}} = \frac{1.04\,\text{pF/cm}}{4.191(10^{-5})\,\text{cm}} = 24.82\,\text{nF/cm}^2$$

The maximum capacitance observed in a high frequency CV plot is just due to the oxide. The minimum capacitance is obtained from a series combination of oxide and depletion capacitances:

$$C_{\min} = \left(\frac{1}{C_{ox}} + \frac{1}{C_s}\right)^{-1} = \left(\frac{1}{68\,\text{nF/cm}^2} + \frac{1}{24.82\,\text{nF/cm}^2}\right)^{-1} = 18.18\ \text{nF/cm}^2$$

Absolute capacitances are just obtained by multiplying by area. The area is just the area of a circular electrode:

$$A = \pi r^2 = (3.1415926\cdots)(50(10^{-4})\,\text{cm})^2 \cong 7.854(10^{-5})\,\text{cm}^2$$

Thus, C_{max} *is* 5.341 pF *and* C_{min} *is* 1.428 pF.

b. Determine flat band capacitance using values obtained in part **a**. Assume that the effective work function difference between pure aluminum and the *p*-type silicon substrate is –0.20 eV. Sketch ideal quasistatic and high frequency CV plots and label minimum, maximum and flat band capacitances and show the position of the flat band voltage.

To calculate flat band capacitance, one must calculate extrinsic Debye length:

$$\lambda_D = \sqrt{\frac{\varepsilon_s kT}{q^2 N_A}} = \sqrt{\frac{(1.04\,\text{pF/cm})(8.6173(10^{-5})\,\text{eV/°K})(300\text{°K})}{(1.602(10^{-19})\,\text{C})(1\,\text{eV/V})(5(10^{15})\,\text{cm}^{-3})}} = 0.05794\,\mu\text{m}$$

It immediately follows that:

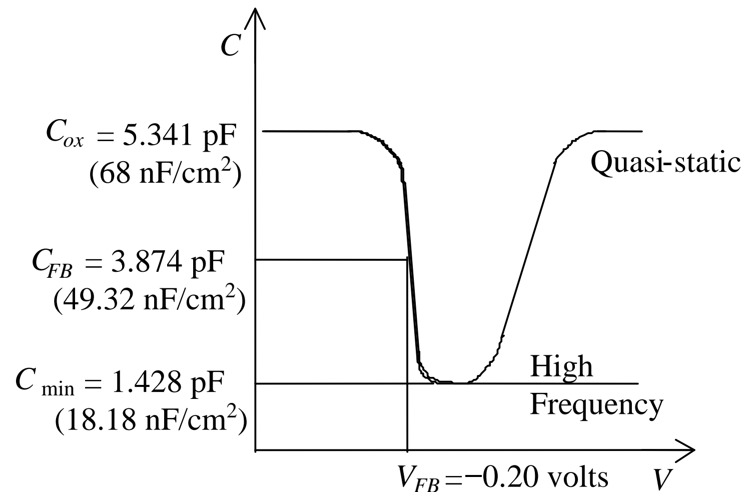
$$C_s^{FB} = \frac{\varepsilon_s}{\lambda_D} = \frac{1.04\,\text{pF/cm}}{5.794(10^{-6})\,\text{cm}} = 179.5\,\text{nF/cm}^2$$

Of course, the measured flat band capacitance is obtained as the series combination of this value with the oxide capacitance:

$$C_{FB} = \left(\frac{1}{C_{ox}} + \frac{1}{C_s^{FB}}\right)^{-1} = \left(\frac{1}{68\,\text{nF/cm}^2} + \frac{1}{179.5\,\text{nF/cm}^2}\right) = 49.32\,\text{nF/cm}^2$$

Obviously, C_{FB} *is* 3.874 pF

Quasi-static and high frequency CV plots can be sketched as follows:



c. How much positive fixed charge per unit area is necessary to cause a flat band shift of 150 mV? If this is due to singly charged interfacial defects, what is the defect density? Does this shift depend on the area of the gate electrode?

The flat band shift due to positive charge must be in the negative direction, thus, the required charge density is:

$$Q_f = -C_{ox}\Delta V_{FB} = -(68\,\text{nF/cm}^2)(-0.15\,\text{V}) = 1.02(10^{-8})\,\text{C/cm}^2$$

If this is due to an interfacial density, N, of singly charged species, this density is:

$$N = \frac{Q_f}{q} = \frac{1.02(10^{-8})\,\text{C/cm}^2}{1.602(10^{-19})\,\text{C}} = 6.365(10^{10})\,\text{cm}^{-2}$$

This is a reasonable result.